Análise de acurácia para algoritmos de classificação

Anderson Gonçalves Marco

***Abstract*—Este trabalho faz uma análise de acurácia dos algoritmos de classificação *IBK*, *KStar*, *naive bayes*, J48, rede neural *multilayer perceptron* e Função Logística, esta análise é realizada sobre grupos de bases de dados com características distintas. As características selecionadas, para as bases, foram a quantidade objetos das bases, número de atributos e número de classes.**

I. Introdução

A mineração de dados é uma área, da ciência da computação, que teve um grande crescimento na ultima década, com varias empresas a empregando largamente. A sua meta é o uso de técnicas para relacionar, visualizar, predizer, e tratar grandes quantidades de dados. Entre os métodos de predição de dados estão os algoritmos de classificação.

Por causa da necessidade de se fazer predição dos dados, hoje existem dezenas de algoritmos de classificação, cada qual com dezenas de variações. A razão para a existência de uma grande quantidade de algoritmos de classificação é que alguns são melhores para determinadas bases de dados do que outros.

Saber que algoritmo deve-se utilizar para que tipo de base de dado não é uma tarefa simples, sendo muitas vezes um processo de escolha ao acaso. Este trabalho tenta fazer um *benchmark*, entre alguns dos principais algoritmos de classificação existentes. Os algoritmos de classificação analisados foram submetidos a grupos de bases de dados, cada qual com características distintas. Assim sendo esperasse conseguir com este trabalho saber que algoritmo de classificação utilizar baseado nas características de uma base de dados.

II. Revisão bibliográfica

Saber que algoritmo de classificação utilizar com base nas características da base de dados é um problema antigo. Sendo algo que remonta a década de 1970.

O artigo de John R. Rice [34] de 1976 foi o primeiro a mostrar a importância deste tipo de problema. Em 1994 houve um ranqueamento dos algoritmos de classificação no livro *Machine Learning, Neural and Statistical Classification* [31]. Em 1998 o artigo de Dmitri A. Rachkovskij e Ernst M. Kussul [33] cria um gerador de bases para testar algoritmos de classificação. Por fim em 2003 o artigo de Brazdil, Pavel B. and Soares, Carlos and Da Costa, Joaquim Pinto [26] faz uma analise de desempenho é acurácia de diferentes algoritmos de classificação.

III. Algoritmos de classificação estudados

Os algoritmos de classificação analisados foram o *IBK*, *KStar*, *naive bayes*, J48, rede neural *multilayer perceptron* e a Função Logística. A seguir uma breve descrição de cada um deles.

- *IBK* [25], algoritmo de classificação baseado no knn, a grande diferença entre ele é o knn é que o *IBK* possui mais parâmetros de ajuste.
- *Kstar* [27], uma variação do knn que usa entropia como medida de distância.
- *Naive Bayes* [29], baseado em probabilidades.
- J48 [32], Algoritmo para geração de arvore de decisão.
- Rede neural *multilayer perceptron* [35], um tipo de rede neural.
- Função Logística [30], algoritmo de classificação que cria planos/hyperplanos lineares.

Todos os algoritmos citados anteriormente são implementados no software Weka [28] versão 3.3.12, com seus parâmetros padrões.

IV. Bases de dados utilizadas

Para a escolha das bases de dados seguiu-se os seguintes critérios:

- Bases com mais de 100 objetos.
- Bases com mais de 10 atributos.

As bases foram dividas em quatro grupos.

- **G1** bases com mais de 2 mil objetos e entre 19 a 40 atributos.
- **G2** Bases que possuem entre 150 a 500 objetos e entre 17 a 23 atributos .
- **G3** Bases que possuem entre 200 a 650 objetos e entre 60 a 170 atributos.
- **G4** Bases que possuem mais do que 6 classes.

As tabela I II, III e IV mostram as bases escolhidas do UCI, junto com algumas características que elas tem, para os grupos **G1**, **G2**, **G3** e **G4** respectivamente:

TABLE I. Bases de dados do grupo **G1**.

nome da base	nº objetos	nº atributos	nº classes
Chees [2],[3]	3196	35	2
Segment [18],[19]	2310	20	7
Sensor [22],[23]	5456	24	4
Wave [24],[11]	5000	40	3

TABLE II. BASES DE DADOS DO GRUPO **G2**.

nome da base	nº objetos	nº atributos	nº classes
Glass [6],[7]	214	10	7
Parkinsons [16],[17]	195	23	2
Heart [12],[13]	270	14	2
Thoraic [20],[21]	470	16	2

TABLE III. BASES DE DADOS DO GRUPO **G3**.

nome da base	nº objetos	nº atributos	nº classes
Hill Valley [8],[9]	606	100	2
Libras [10],[1]	606	91	15
Sonar [4],[5]	208	60	2
Musk [14],[15]	476	168	2

TABLE IV. BASES DE DADOS DO GRUPO **G4**.

nome da base	nº objetos	nº atributos	nº classes
Segment [18],[19]	2310	20	7
Glass [6],[7]	214	10	7
Libras [10],[1]	606	91	15

V. APLICAÇÃO DOS ALGORITMOS DE CLASSIFICAÇÃO NAS BASES DE DADOS

A aplicação dos algoritmos foi realizada utilizando-se o método de k-pastas, onde cada pasta corresponde a 10% dos elementos da base, este método foi repetido 10 vezes, totalizando assim 100 aplicações de um determinado algoritmo para uma determinada base.

As tabelas V, VI, VII e VIII mostram a média da porcentagem de acerto, ou acurácia, que cada um dos algoritmos obteve para as diferentes bases. As tabelas IX, X, XI e XII mostram o desvio padrão referente as médias existentes das tabelas V, VI, VII e VIII.

TABLE V. MÉDIA DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G1**.

	Chess [2]	Segment [18]	Sensor [22]	Wave [24]
***IBK* [25]**	96.13	97.15	88.43	73.41
***Kstar* [27]**	96.91	97.09	94.68	73.36
Bayes [29]	87.79	80.11	52.50	80.01
J48 [32]	99.44	96.74	99.61	75.25
Perceptron [35]	99.31	96.27	88.04	83.56
Logística [30]	97.56	95.61	70.42	86.73

TABLE VI. MÉDIA DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G2**.

	Glass [6]	Parkinsons [16]	Heart [12]	Thoraic [20]
***IBK* [25]**	88.93	95.91	76.15	77.02
***Kstar* [27]**	91.13	89.12	76.44	82.23
Bayes [29]	83.13	70.14	83.59	77.74
J48 [32]	97.33	84.74	78.15	84.64
Perceptron [35]	95.66	91.08	91.08	80.91
Logística [30]	93.93	85.94	85.94	82.77

TABLE VII. MÉDIA DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G3**.

	Hill Valley [8]	Libras [10]	Sonar [4]	Musk [14]
***IBK* [25]**	55.90	86.06	86.17	93.34
***Kstar* [27]**	53.09	83.78	85.11	72.45
Bayes [29]	51.93	64.14	67.71	76.16
J48 [32]	50.66	69.36	73.61	99.22
Perceptron [35]	56.19	80.39	81.61	57.29
Logística [30]	81.65	68.97	72.47	98.26

TABLE VIII. MÉDIA DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G4**.

	Segment [18]	Glass [6]	Libras [10]
***IBK* [25]**	97.15	88.93	86.17
***Kstar* [27]**	97.09	89.12	85.11
Bayes [29]	80.11	70.14	67.71
J48 [32]	96.74	84.74	73.61
Perceptron [35]	96.27	91.08	81.61
Logística [30]	95.61	85.94	72.47

TABLE IX. DESVIO PADRÃO DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G1**.

	Chess [2]	Segment [18]	Sensor [22]	Wave [24]
***IBK* [25]**	1.12	1.11	1.30	1.82
***Kstar* [27]**	0.92	1.12	0.89	1.93
Bayes [29]	1.91	2.11	1.96	1.45
J48 [32]	0.37	1.28	0.31	1.90
Perceptron [35]	0.44	1.25	2.10	1.61
Logística [30]	0.66	1.52	1.49	1.49

TABLE X. DESVIO PADRÃO DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G2**.

	Glass [6]	Parkinsons [16]	Heart [12]	Thoraic [20]
***IBK* [25]**	5.21	4.52	8.46	5.06
***Kstar* [27]**	5.44	7.56	7.42	2.76
Bayes [29]	7.59	9.30	5.98	9.44
J48 [32]	3.35	8.01	7.42	1.15
Perceptron [35]	4.10	6.71	7.09	4.04
Logística [30]	5.13	7.79	6.43	2.68

TABLE XI. DESVIO PADRÃO DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G3**.

	Hill Valley [8]	Libras [10]	Sonar [4]	Musk [14]
***IBK* [25]**	6.51	5.21	8.45	3.24
***Kstar* [27]**	6.07	5.28	7.65	6.56
Bayes [29]	4.12	7.90	8.66	7.76
J48 [32]	0.50	8.38	9.34	1.70
Perceptron [35]	1.76	5.89	8.66	4.27
Logistic[30]	5.54	7.96	8.90	2.35

TABLE XII. DESVIO PADRÃO DA PORCENTAGEM DOS ACERTOS DAS BASES DO GRUPO **G4**.

	Segment [18]	Glass [6]	Libras [10]
***IBK* [25]**	1.11	5.21	5.21
***Kstar* [27]**	1.12	5.44	5.28
Bayes [29]	2.11	7.59	7.90
J48 [32]	1.28	3.35	8.38
Perceptron [35]	1.25	4.10	5.89
Logística [30]	1.52	5.13	7.96

As tabelas XIII, XIV, XV e XVI têm a média das médias de acertos dos algoritmos de classificação e também têm uma

média do desvio padrão que o algoritmos de classificação obtiveram para as bases dos grupos **G1**, **G2**, **G3**, **G4**.

TABLE XIII. MÉDIA DAS MÉDIAS DOS ACERTOS E DO DESVIO PADRÃO DAS BASES DO GRUPO **G1**.

Algoritmo	Média das médias	Média do desvio padrão
***IBK* [25]**	88.78	1.34
***Kstar* [27]**	90.51	1.22
Bayes [29]	75.10	1.86
J48 [32]	92.76	0.96
Perceptron [35]	91.80	1.30
Logística [30]	87.58	1.45

TABLE XIV. MÉDIA DAS MÉDIAS DOS ACERTOS E DO DESVIO PADRÃO DAS BASES DO GRUPO **G2**.

Algoritmo	Média das médias	Média do desvio padrão
***IBK* [25]**	84.50	5.81
***Kstar* [27]**	84.73	5.79
Bayes [29]	78.65	8.08
J48 [32]	86.21	4.98
Perceptron [35]	86.76	5.48
Logística [30]	86.58	5.51

TABLE XV. MÉDIA DAS MÉDIAS DOS ACERTOS E DO DESVIO PADRÃO DAS BASES DO GRUPO **G3**.

Algoritmo	Média das médias	Média do desvio padrão
***IBK* [25]**	80.37	5.85
***Kstar* [27]**	73.61	6.39
Bayes [29]	64.98	7.11
J48 [32]	73.21	4.98
Perceptron [35]	68.87	5.14
Logística [30]	80.34	6.19

TABLE XVI. MÉDIA DAS MÉDIAS DOS ACERTOS E DO DESVIO PADRÃO DAS BASES DO GRUPO **G4**.

Algoritmo	Média das médias	Média do desvio padrão
***IBK* [25]**	90.71	3.84
***Kstar* [27]**	90.66	3.95
Bayes [29]	75.79	5.87
J48 [32]	87.81	4.34
Perceptron [35]	90.77	3.75
Logística [30]	86.17	4.87

Um fado interessante é o baixo valor do atributo **Média do desvio padrão** na tabela XIII em comparação com os das tabelas XIV e XV. O autor acredita que isto acontece por que esta tabela é referente ao grupo **G1**. O grupo **G1** tinha como característica todas as suas bases terem muitos objetos, isto pode ter pode possibilitado com que houvesse uma convergência maior entre os diferentes modelos gerados para uma determinada base. Esta hipótese é reforçada pelo fato da tabela IX também apresentar valores mais baixos de desvio padrão padrão dos que as tabelas X e XI.

VI. ANÁLISE DA ACURÁCIA

Para esta análise foi determinado que o melhor classificador, para um determinado grupo de base de dados, é aquele que possui a maior média das médias de acertos presentes nas tabelas XIII, XIV, XV e XVI . Também foi determinado que as tolerâncias para a acurácia são as maiores médias do desvio padrão das tabelas XIII, XIV, XV e XVI. Assim a tolerância para o grupo **G1** é de 1.86%, para o grupo **G2** é de 8.08%, para o grupo **G3** é de 7.11% e para o grupo **G4** é 5.87%.

A análise de acurácia então indica que para as bases do grupo **G1** existe um empate, para o melhor classificador, entre o Perceptron [35] é o Logística [30]. Para as bases do grupo **G2** houve um empate entre todos classificadores. Para as bases do grupo **G3** o *IBK* [25], o Logística [30] e o *Kstar* [27] ficam empatados, entretanto pode-se observar que o *Kstar* [27] quase não conseguiu ficar entre os melhores classificadores para este grupo. Para as bases do grupo **G4** todos os classificadores, com exceção do Bayes [29] que ficou abaixo dos outros, ficam empatados.

Uma observação importante é que o Bayes [29] possui sempre o ultimo lugar nas tabelas XIII, XIV, XV e XVI. Este fato indica que ele esta para os algoritmos de classificação (em termos de acurácia) como o algoritmo *Bubble Sort* esta (em termos de desempenho) para os algoritmos de ordenação.

VII. CONCLUSÃO

Neste trabalho foi realizado uma análise de acurácia entre alguns dos algoritmos de classificação mais clássicos, esta análise foi feita em grupos de bases de dados com características distintas. A análise mostrou que para bases com muitos objetos os algoritmos Função Logística [30] e rede neural multilayer perceptron [35] são os melhores. Para bases com muitos atributos *IBK* [25], Função Logística [30] e *Kstar* [27] são os melhores. A análise também concluiu que o classificador *naive bayes* [29] foi oque apresentou os piores resultados, para os grupos de bases analisadas.

REFERENCES

[1] Libras Movement Data Set , howpublished = http://archive.ics.uci.edu/ml/machine-learning-databases/libras/movement_libras.data, note = Accessed: 2015-07-30.

[2] Chess (king-rook vs. king-pawn) data set. https://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King-Pawn%29. Accessed: 2015-07-30.

[3] Chess (king-rook vs. king-pawn) data set. https://archive.ics.uci.edu/ml/machine-learning-databases/chess/king-rook-vs-king-pawn/kr-vs-kp.data. Accessed: 2015-07-30.

[4] Connectionist bench (sonar, mines vs. rocks) data set. https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+%28Sonar%2C+Mines+vs.+Rocks%29. Accessed: 2015-07-30.

[5] Connectionist bench (sonar, mines vs. rocks) data set. https://archive.ics.uci.edu/ml/machine-learning-databases/undocumented/connectionist-bench/sonar/sonar.all-data. Accessed: 2015-07-30.

[6] Glass identification data set. https://archive.ics.uci.edu/ml/datasets/Glass+Identification. Accessed: 2015-07-30.

[7] Glass identification data set. https://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.data. Accessed: 2015-07-30.

[8] Hill-valley data set. https://archive.ics.uci.edu/ml/datasets/Hill-Valley. Accessed: 2015-07-30.

[9] Hill-valley data set. http://archive.ics.uci.edu/ml/machine-learning-databases/hill-valley/Hill_Valley_with_noise_Testing.data. Accessed: 2015-07-30.

[10] Libras movement data set. https://archive.ics.uci.edu/ml/datasets/Libras+Movement. Accessed: 2015-07-30.

[11] MS Windows NT kernel description. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2015-07-30.

[12] MS Windows NT kernel description. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2015-07-30.

[13] MS Windows NT kernel description. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2015-07-30.

[14] Musk (version 1) data set. https://archive.ics.uci.edu/ml/datasets/Musk+%28Version+1%29. Accessed: 2015-07-30.

[15] Musk (version 1) data set. https://archive.ics.uci.edu/ml/machine-learning-databases/musk/clean1.data.Z. Accessed: 2015-07-30.

[16] Parkinsons data set. https://archive.ics.uci.edu/ml/datasets/Parkinsons. Accessed: 2015-07-30.

[17] Parkinsons data set. https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data. Accessed: 2015-07-30.

[18] Statlog (image segmentation) data set. https://archive.ics.uci.edu/ml/datasets/Statlog+%28Image+Segmentation%29. Accessed: 2015-07-30.

[19] Statlog (image segmentation) data set. https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/segment/segment.dat. Accessed: 2015-07-30.

[20] Thoracic surgery data data set. https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data. Accessed: 2015-07-30.

[21] Thoracic surgery data data set. https://archive.ics.uci.edu/ml/machine-learning-databases/00277/ThoracicSurgery.arff. Accessed: 2015-07-30.

[22] Wall-following robot navigation data data set. https://archive.ics.uci.edu/ml/datasets/Wall-Following+Robot+Navigation+Data. Accessed: 2015-07-30.

[23] Wall-following robot navigation data data set. https://archive.ics.uci.edu/ml/machine-learning-databases/00194/sensor_readings_24.data. Accessed: 2015-07-30.

[24] Wall-following robot navigation data data set. http://web.archive.org/web/20080207010024/http://www.808multimedia.com/winnt/kernel.htm. Accessed: 2015-07-30.

[25] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[26] Pavel B. Brazdil, Carlos Soares, and Joaquim Pinto Da Costa. Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. *Mach. Learn.*, 50(3):251–277, March 2003.

[27] John G. Cleary and Leonard E. Trigg. K*: An instance-based learner using an entropic distance measure. In *12th International Conference on Machine Learning*, pages 108–114, 1995.

[28] G. Holmes, A. Donkin, and Ian H. Witten. Weka: a machine learning workbench. In *Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*, pages 357–361, Nov 1994.

[29] George H. John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[30] S. le Cessie and J.C. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[31] Donald Michie, D. J. Spiegelhalter, C. C. Taylor, and John Campbell, editors. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Upper Saddle River, NJ, USA, 1994.

[32] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[33] Dmitri A. Rachkovskij and Ernst M. Kussul. Datagen: a generator of datasets for evaluation of classification algorithms. *Pattern Recognition Letters*, 19(7):537–544, 1998.

[34] John R. Rice. The algorithm selection problem. *Advances in Computers*, (15):65–118, 1976.

[35] Frank Rosenblatt. *Principles of Neurodynamics*. Spartan Books, New York, 1962.